**Baseball Summary Analysis**

Team 6:

Michael Good

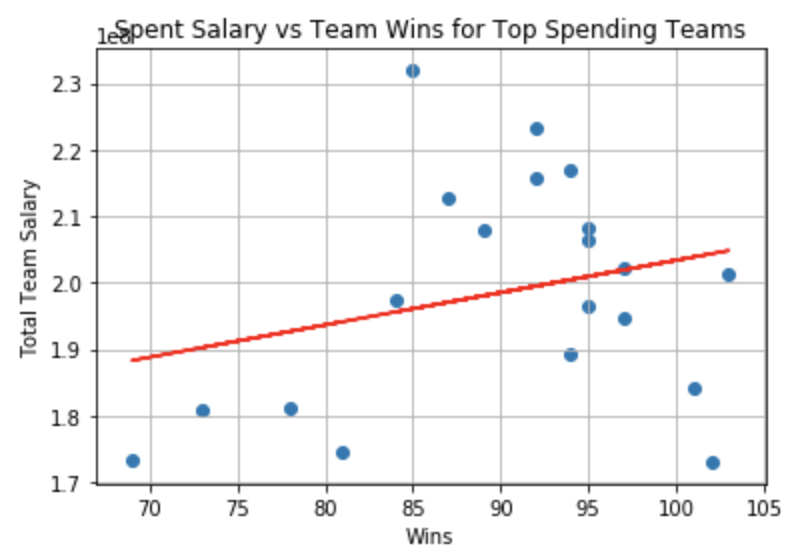
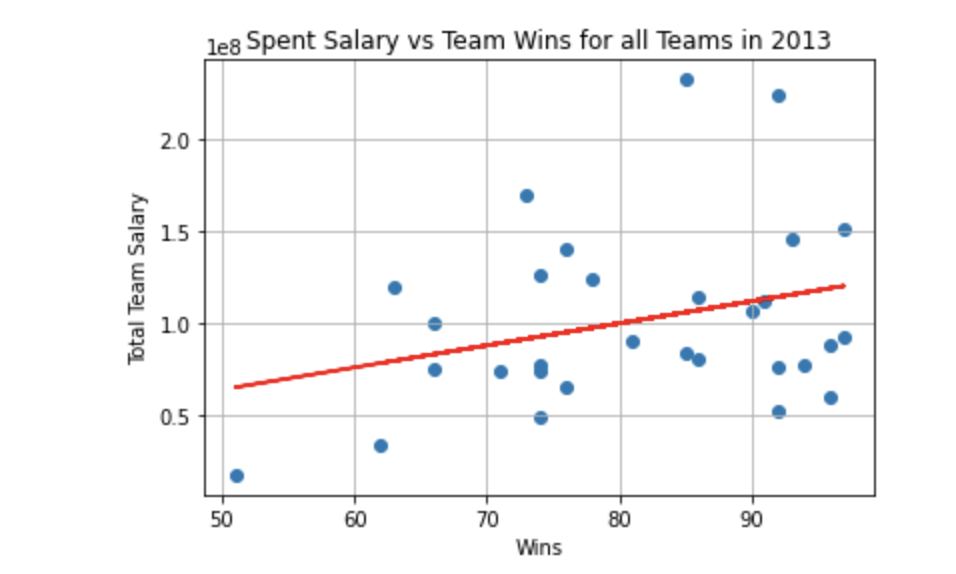
Kimberly Kockenmeister

Amanda Vital

Ashlin Spence

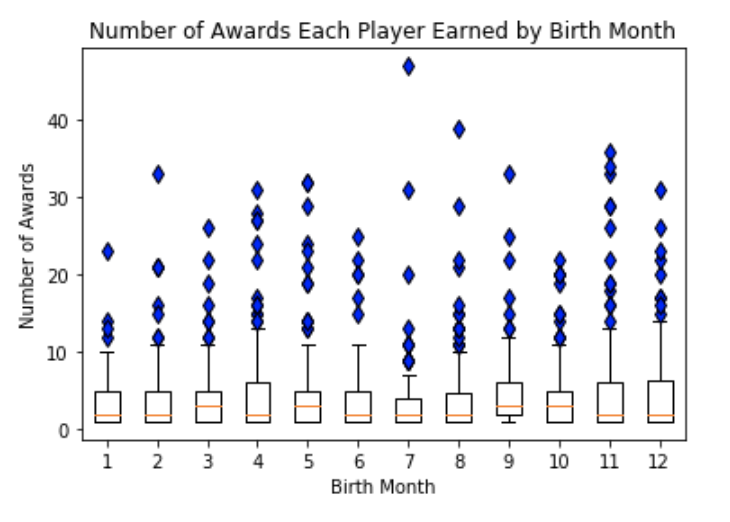
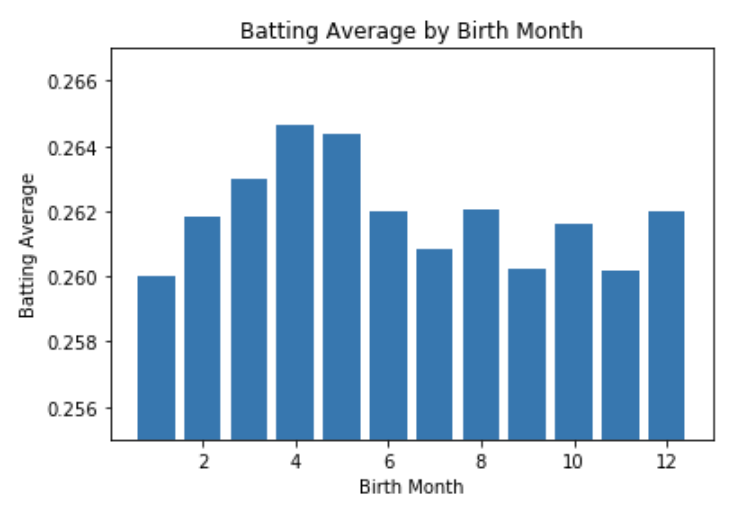
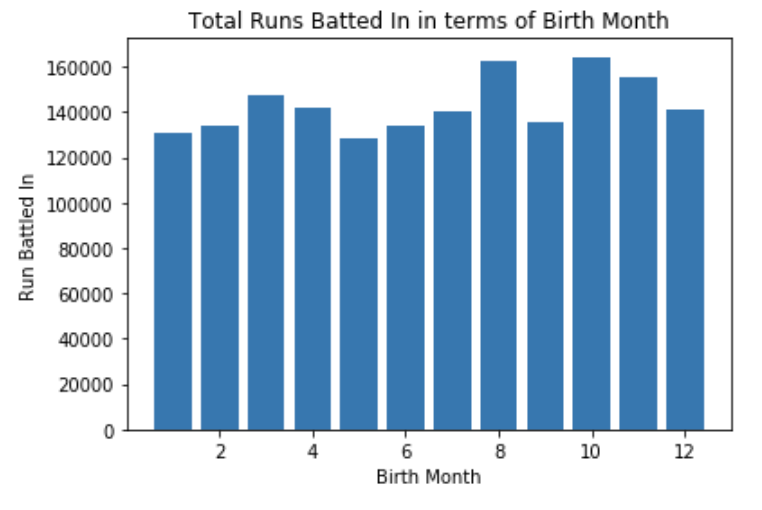
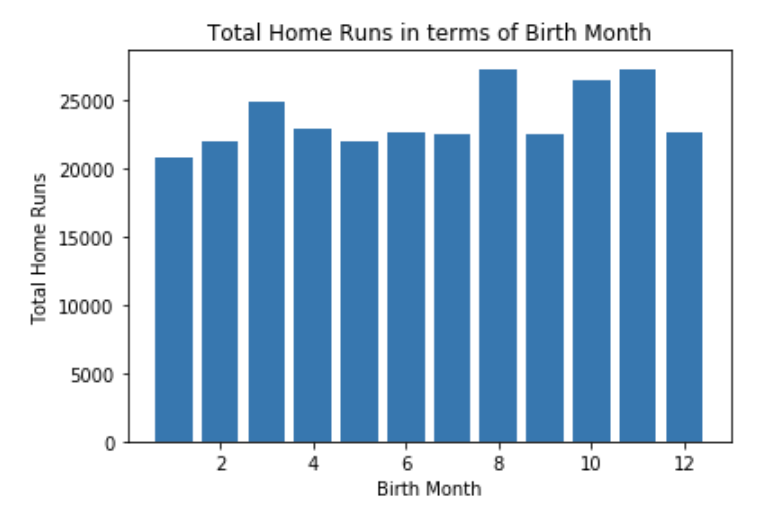
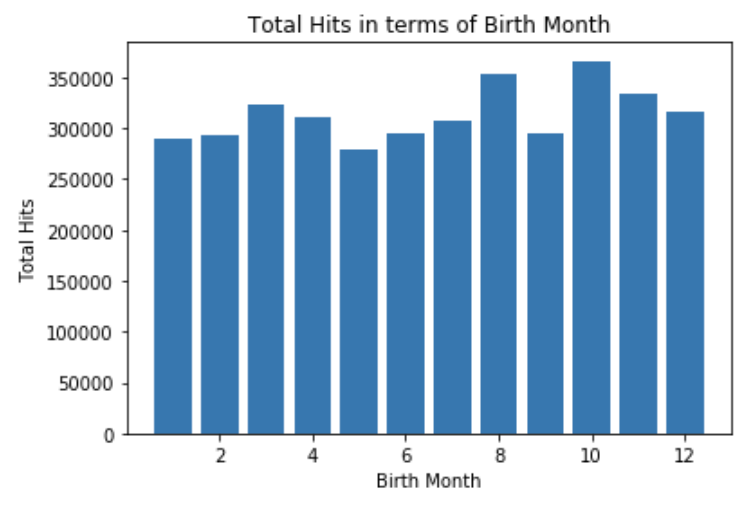
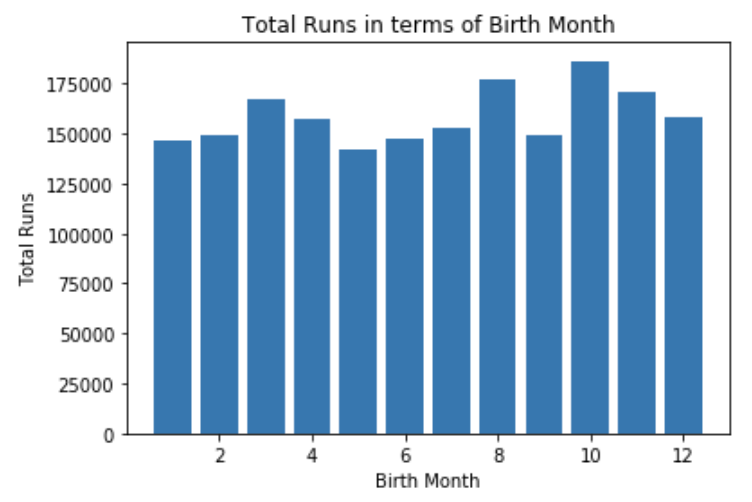
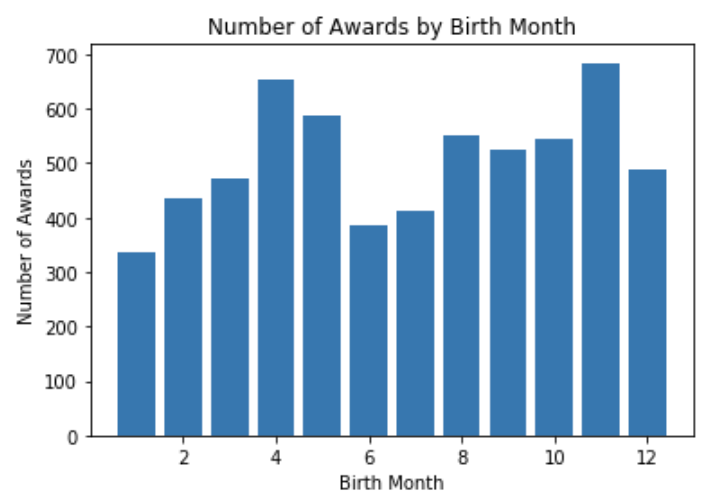
Jaimie Koshy

**Salary and Team Record**

1. Findings-
   1. There is no relationship between money spent on salaries and team wins
2. Graphs -
   1. 
3. Analysis-
   1. We took two csv files and created two different data frames so that we could pull the information we needed. Then we merged these data frames together so we could look at the bigger picture. We utilized different functions such as groupby, sum, sort values, and linregress to manipulate our data. We then graphed this using scatter plots so we could try to draw a conclusion between the two variables. We saw that there was no correlation between money spent on salaries and team wins because we found the r value was both less than .3 which shows since it is close to 0 there is no relationship.

**Birth Month and Player Skill**

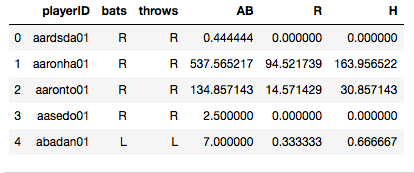
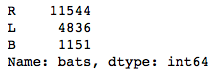
We wanted to analyze whether the month a player was born in had a clear effect on his performance in the MLB. We plotted birth month according to a number of different statistics and obtained the following charts:



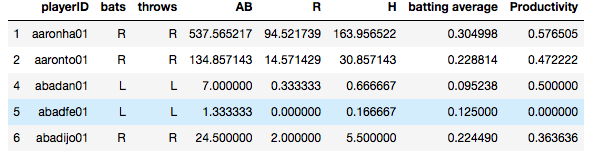
These charts demonstrate that there is very little correlation between birth month and any batting related statistic. We also see that for all the players here the average number of hits between runs and homeruns was roughly equivalent because of the very similar looking graphs. The only reason the batting average appears to have an effect is because its y-axis was very zoomed in. In reality there is no correlation between birth month and batting average. Finally, we plotted the number of awards each player achieved and created a box plot organized by birth month. As expected the values are clustered around 1 award with numerous upper outliers all the way up to Barry Bonds born in July with 47 awards.

**Handedness vs. Batting Average**

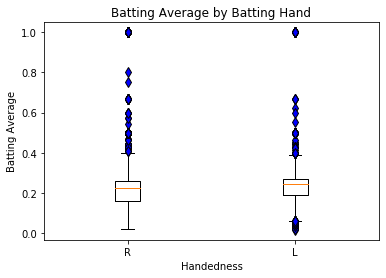
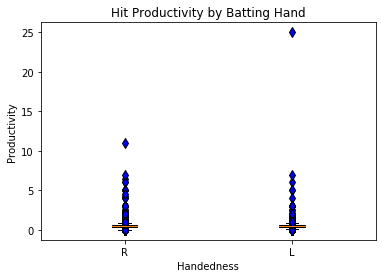
We wanted to examine whether or not handedness had an affect on Batting average. In order to do this we first examined both data set’s separately and then combined them into one DataFrame. I think it’s important to add that the csv files we used for this question, as well as many of the other questions, consisted of data from 1877 - 2015. After combining the two DataFrames, we decided to clean it up. There were a lot of “Nan” values so we removed them before continuing our analysis of the DataFrames. Below I've inserted images from our jupyter notebook to show how we counted which hand players use to bat and throw, I have also included the combined DataFrame.



After combining the DataFrames we calculated the batting average and the productivity for each ‘playerID’. To calculate the batting average, we divided ‘H’(Hits) by ‘AB’(At-Bats). To calculate production, we divided ‘R’(Runs) by ‘H’(Hits).



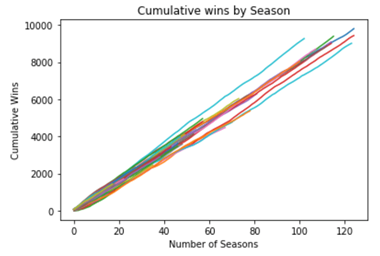
Once we had everything we needed, we continued to our next step which is where we used a boxplot to model the relationship between batting average and handedness as well as productivity and handedness.



We performed a t-test for both. For productivity vs Handedness we found that our p-value wass slightly above our critical threshold of 0.05. This means that there does not appear to be a statistically significant difference between the two different populations. For Batting Average vs Handedness our p-value was very low which means that we reject our null hypothesis and that our results are statistically significant. In the end, our results showed that Left Handed batters have a statistically significantly higher batting average than the Right Handed batters.

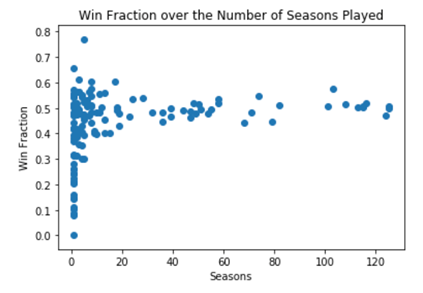
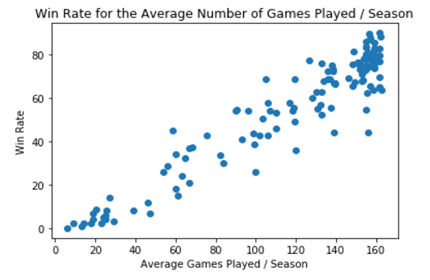
**Most Successful Franchise**

We set out to evaluate which team was the most successful franchise in the history of the MLB. First, we considered that the team with the most overall wins over time would be the most successful team. However, this would bias the analysis towards teams that had been in the league for a long time. A team may be around for a long time and be completely mediocre. We next graphed the cumulative wins of all teams in our dataset who were in the league for greater than 15 seasons.



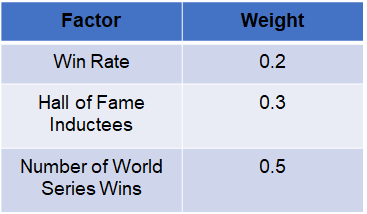
The most interesting factor in this plot is not the achieved y value by teams, but rather the slope of the lines in the plot. It is clear that there is one team in particular that has a significantly higher slope than the rest of the teams.

However, it is possible that this win rate is not as indicative of teams success as we first thought. The following two plots both indicate some confounding factors in this analysis.

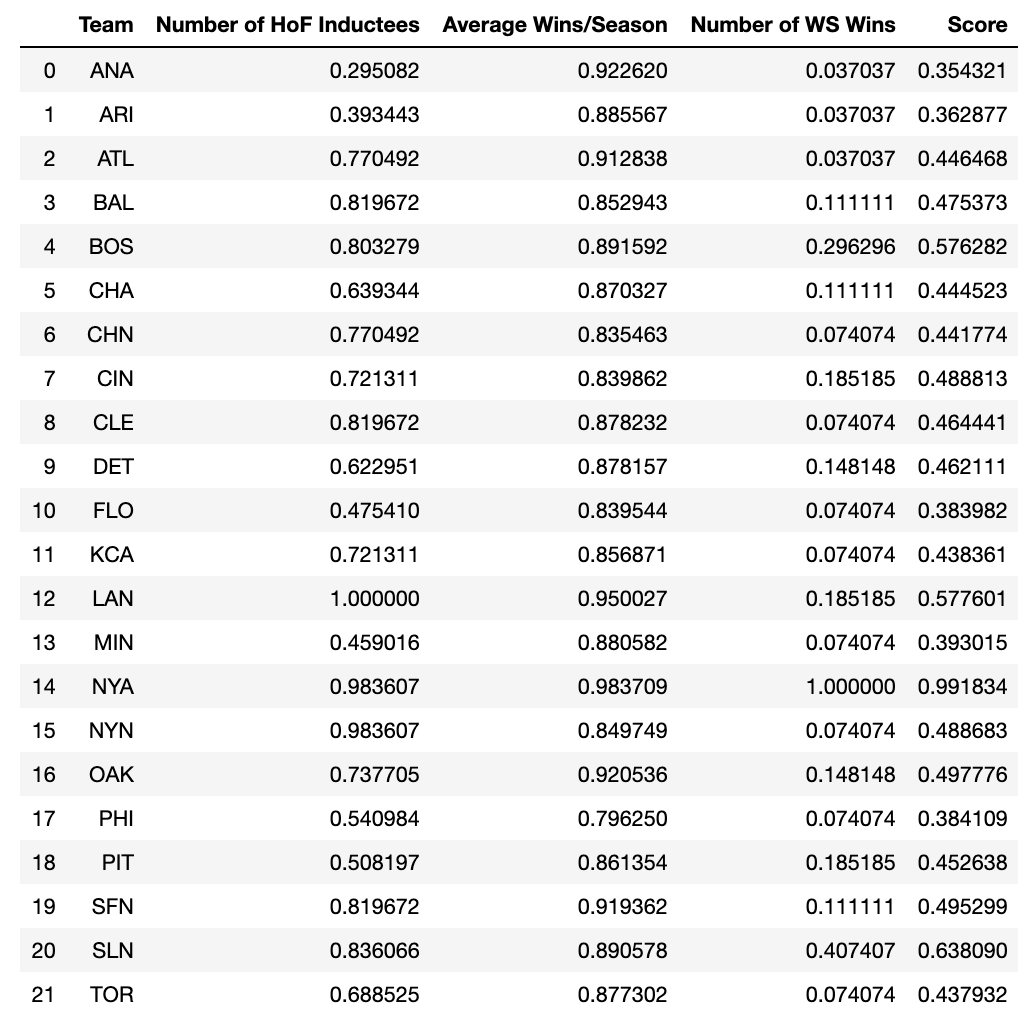


As we can see from both of these plots, the more games a team plays in a season the higher their win rate is. This biases us significantly towards teams that have been playing in more recent years when seasons have become longer. We also see that teams that are in the league for over 10 years tend to gravitate towards a 0.500 record. This indicates that while a team may have a good season or a few, it is averaged out by worse seasons. So a higher win rate than the rest of teams may be an important element in success but there are clearly others.

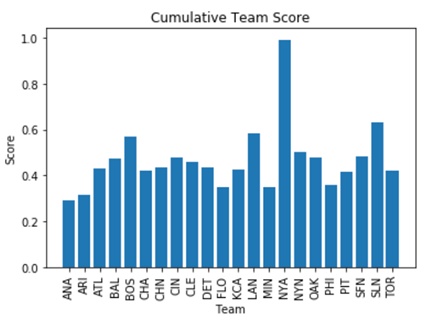
As a result, we have created a model to account for different factors of team success.



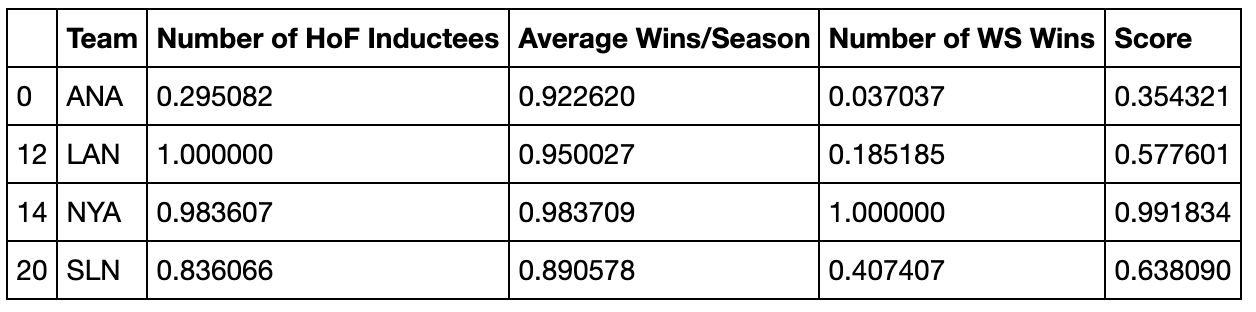
For each factor we assign a wait for relative importance. Win rate, discussed above, is important over time, but does not tell the whole story. The number of Hall of Fame Inductees indicates the level of star power that a team has had. In a team sport a Hall of Fame Inductee is supported by the rest of the team. A pitcher may be very successful, but if the rest of the team constantly drops fly balls then he is unlikely to be a Hall of Fame pitcher. Finally, “World Series Wins” is the most heavily weighted factor. This factor takes into account both the success of a team on the franchise level and also the success from the perspective of the fans. A team may be generally successful, but if they reach the World Series every year only to lose then the fans are unlikely to view this team as a success. This allows us to create the following table and plot:



This table contains the normalized data and the score for each team. The data was normalized based upon the maximum value in the column. Unfortunately, due to the limitations of the dataset the team ID code was not consistent across the data set. These 21 teams were the only ones who had data in each of the above listed categories.



As can be clearly seen, there is one team that is a clear outlier from the rest of the dataset. However, to be more rigorous we calculated the IQR of the dataset and determined outliers.



This table shows the outliers of the dataset. As can be clearly seen, one team, the New York Yankees, achieved a score of 0.991834 out of a possible 1.00. The next highest team is the St. Louis Cardinals with a score of 0.638090. From this analysis we can clearly see that the most successful franchise of all time in baseball is the New York Yankees.

One caveat of this analysis is that, given the high weighting of World Series Wins, the fact that the Yankees have won more world series’ than the next 4 teams combined clearly contributed to their domination of this chart.